

Data Mining Process

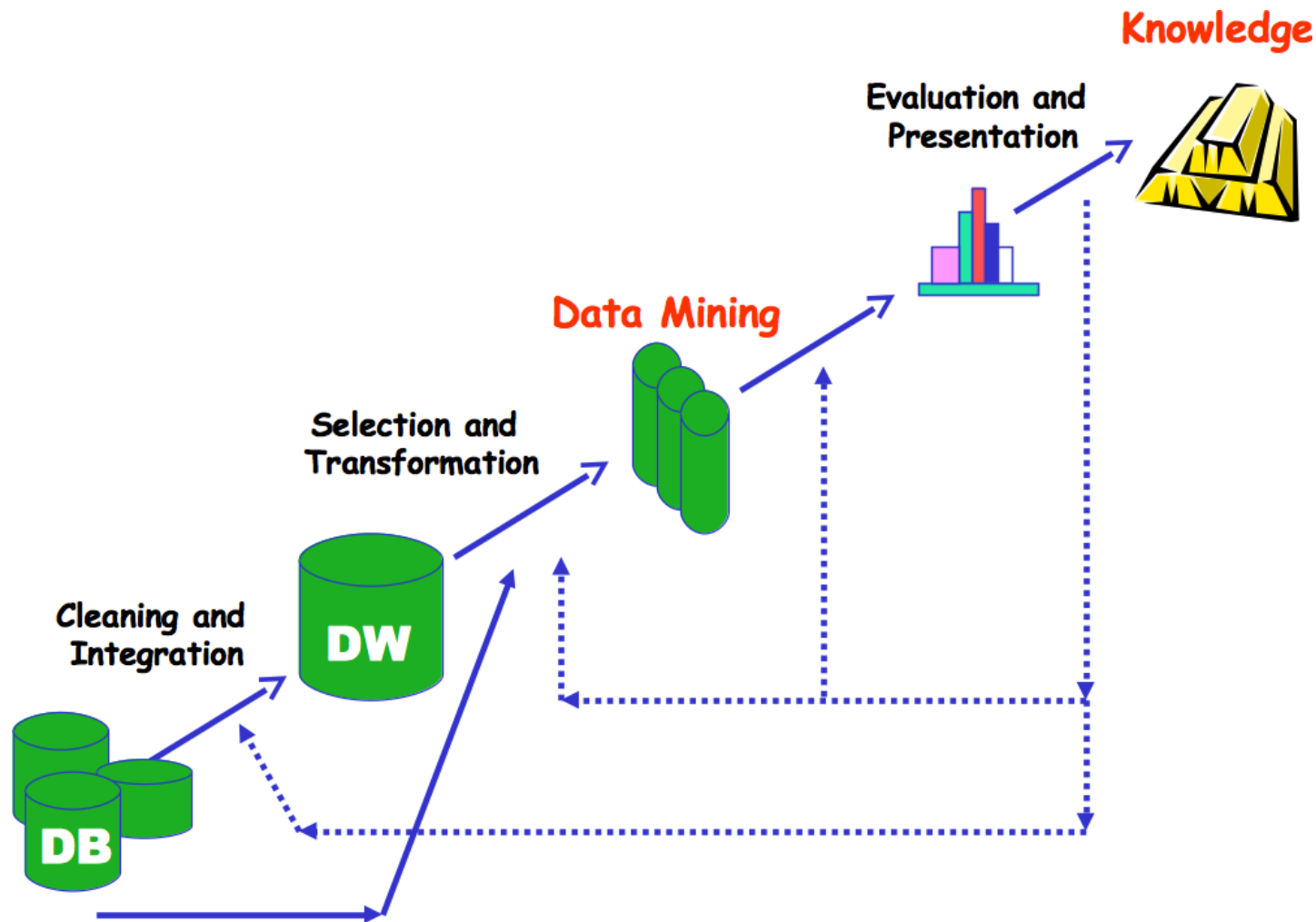
Carlos Soares

(partly using materials kindly
provided by José Luís Borges)

reminder: data mining?

- (or Knowledge Discovery in Databases)
- Is the non-trivial process of identifying
 - implicit (by contrast to explicit)
 - valid (patterns should be valid on new data)
 - novel (novelty can be measured by comparing to expected values)
 - potentially useful (should lead to useful actions)
 - understandable (to humans)
- patterns in data
- Data Mining
 - is a step in the KDD process
 - (arguable, but who cares anyway!...)

the KDD process



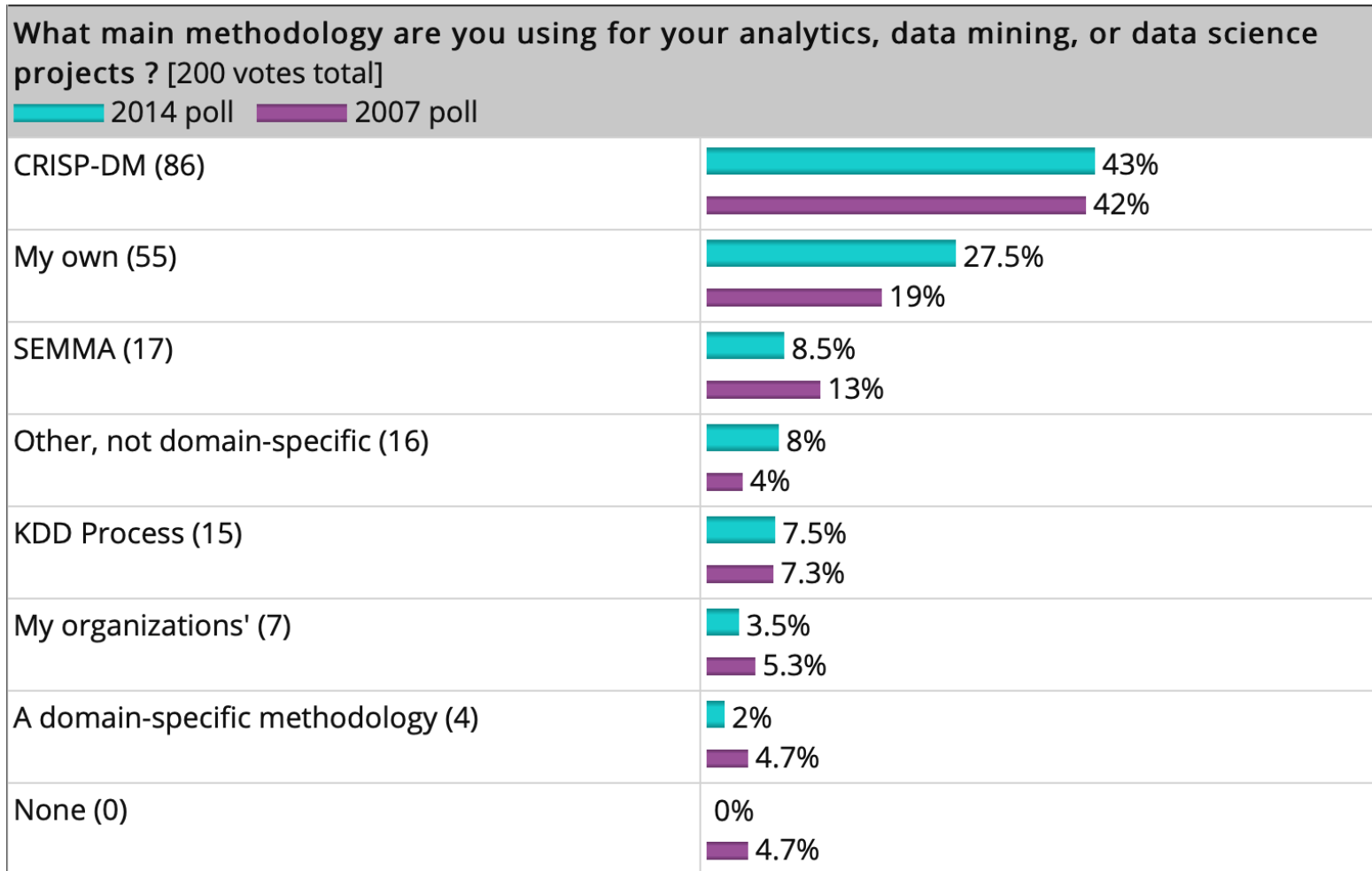
wanted: DM methodology

- Framework for recording experience
 - Allows projects to be replicated
- Aid to project planning and management
- “Comfort factor” for new adopters
 - Demonstrates maturity of Data Mining
 - Reduces dependency on “stars”
- Encourage best practices and help to obtain better results

plan

- methodologies
 - CRISP-DM
 - SCRUM-DM
- ... and beyond
 - MLOps

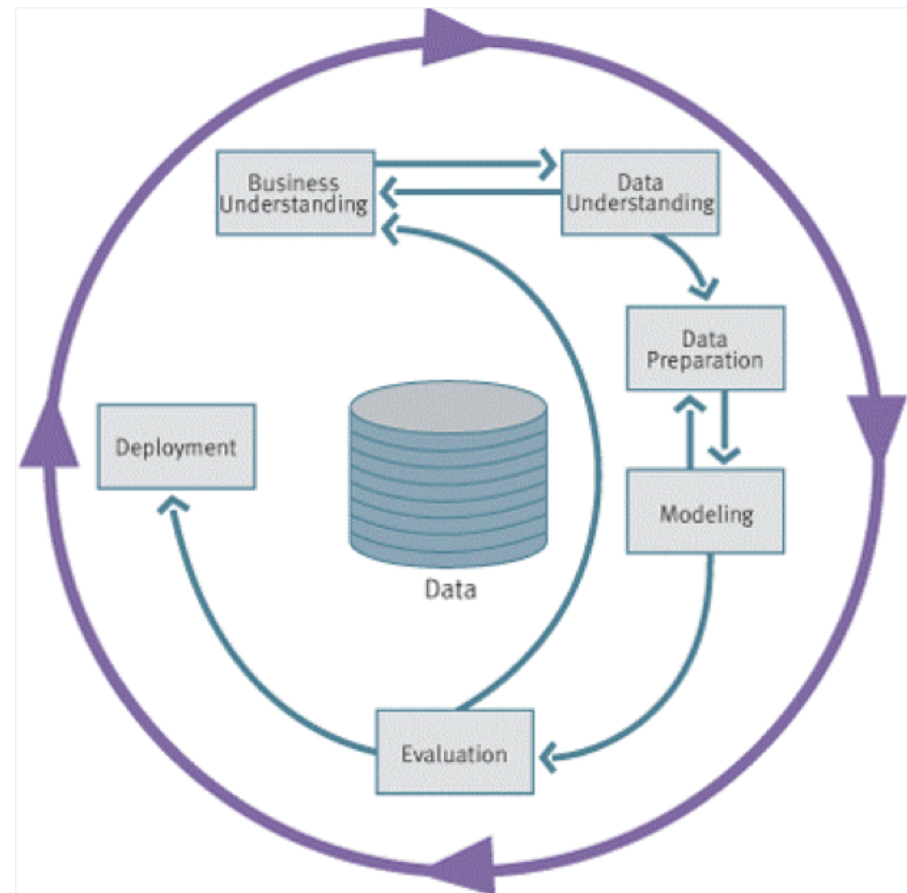
why CRISP-DM?



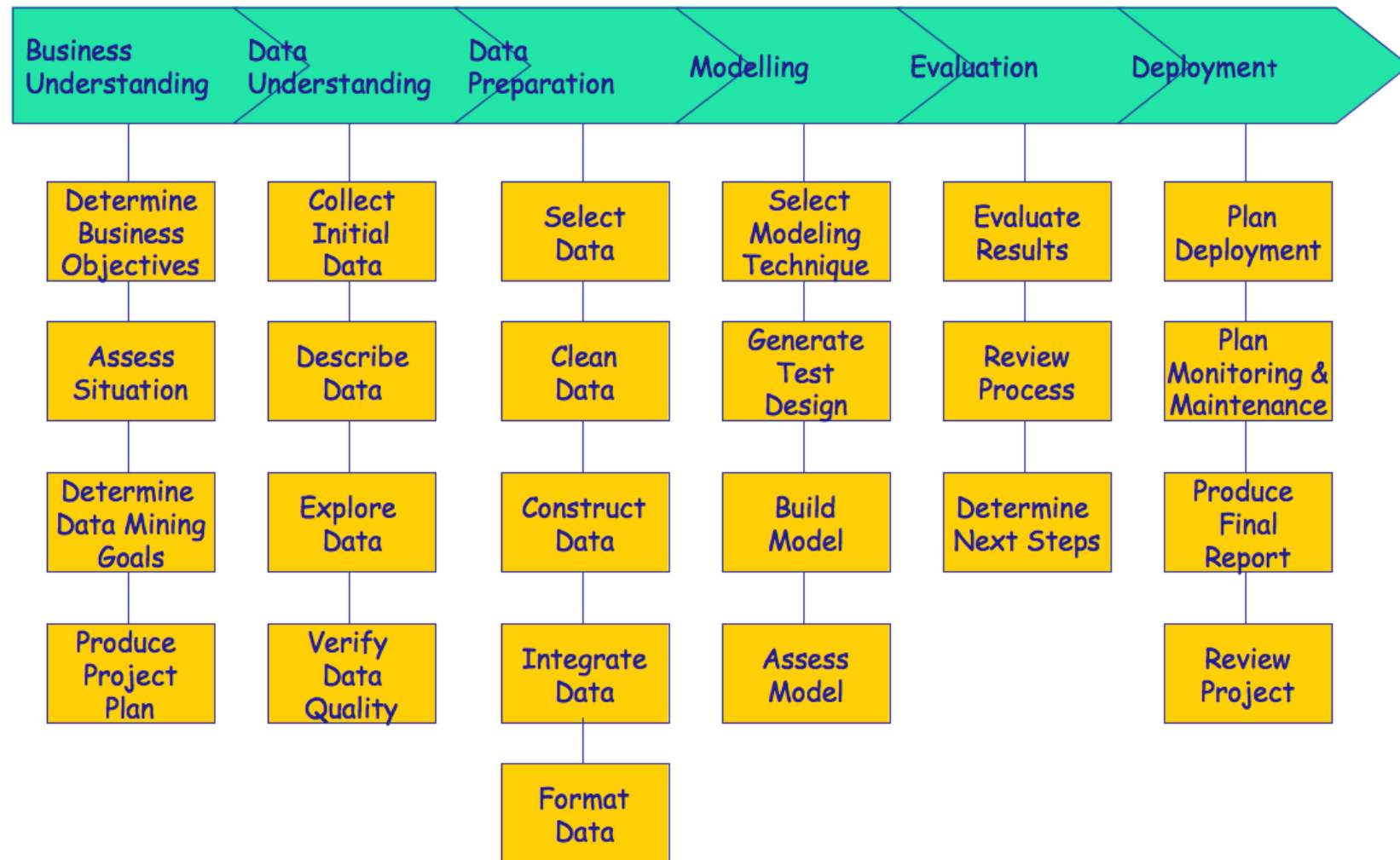
<https://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html>

CRISP-DM: overview

- cross-Industry Standard Process for Data Mining (CRISP-DM)
- European Community funded effort aiming to
 - cheaper, faster, and more reliable data mining
 - widespread adoption
 - reduce skills required for data mining
 - capture experience for reuse
- characteristics
 - non-proprietary
 - application/industry neutral
 - tool neutral
 - focus on business issues
 - as well as technical analysis
 - framework for guidance
 - experience based
 - templates for analysis



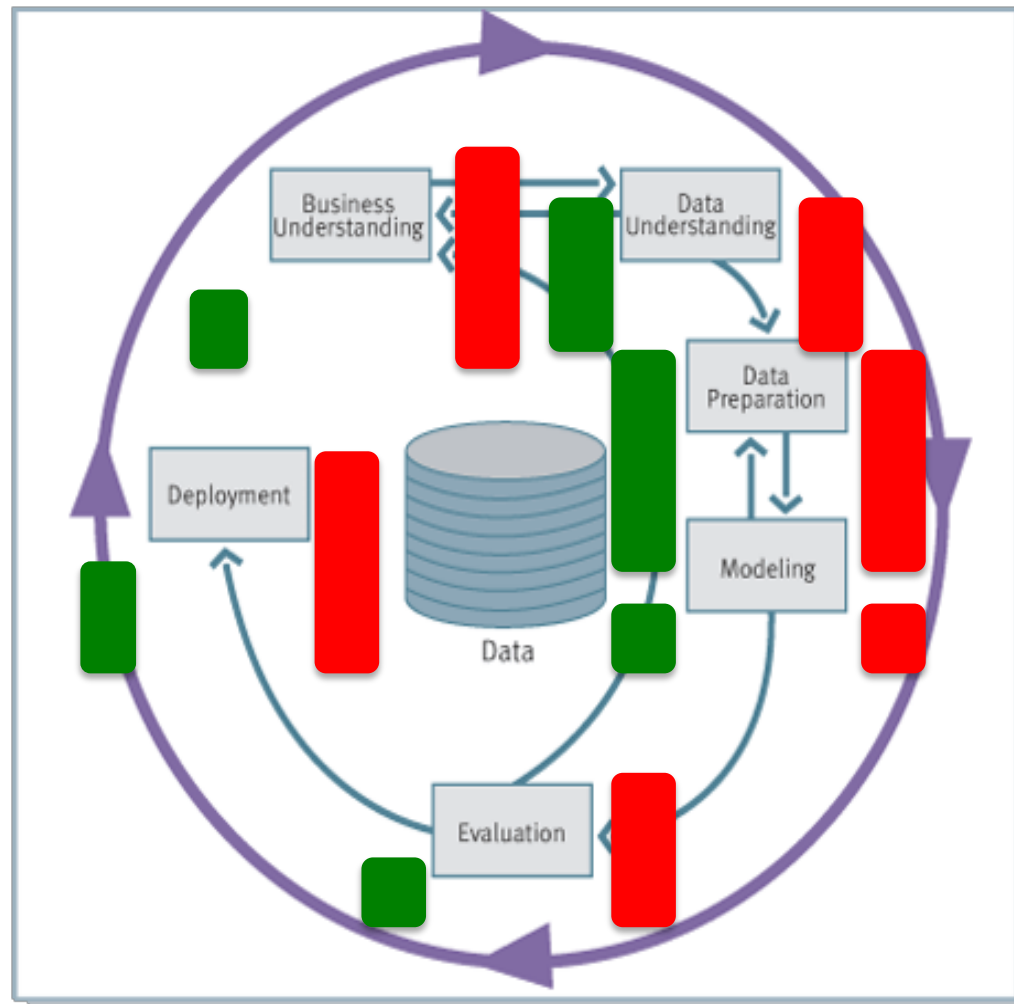
CRISP-DM: phases and tasks



other methodologies

- SEMMA
 - <http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html>
 - SAS Enterprise Miner
- Others
 - specific
 - <http://datalligence.blogspot.com/2008/12/data-mining-methodologies.html>
- Essentially equivalent

effort/impact on success



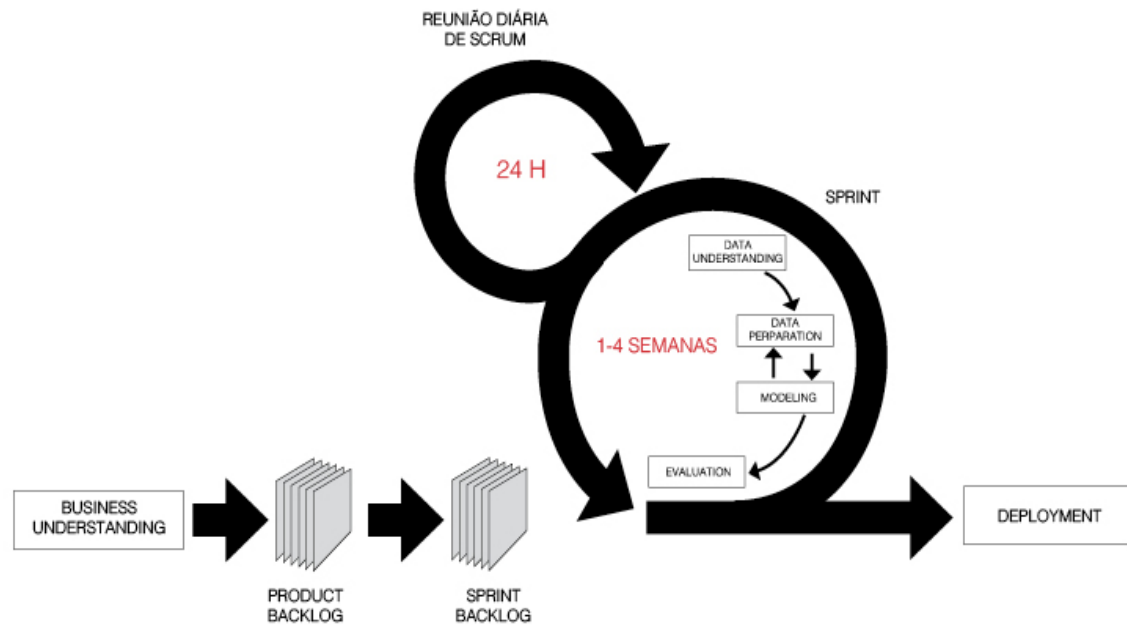
SCRUM-DM: an agile DM methodology

Scrum
work management



CRISP-DM
Data Mining

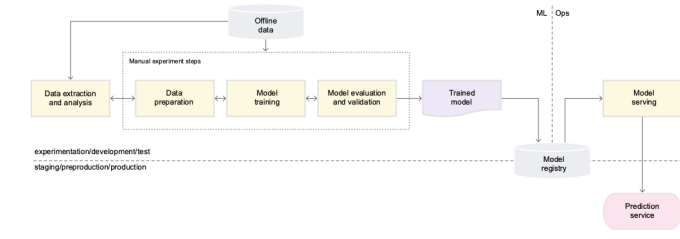
- 3 phases
 - Business Understanding
 - Sprint
 - Deployment
- 6 concepts
 - Product Owner
 - Scrum Master
 - Development team
 - Data Mining Story
 - Product Backlog
 - Sprint Backlog



M.Sc. of Diogo Nogueira, joint work with Ana Barros, Paula Gomes and Ademar Aguiar

AI/ML is not (only) software engineering

- versioning
 - data
 - model
- automation
 - development
 - testing
- collaborative model development
- deployment
 - monitoring
 - maintenance



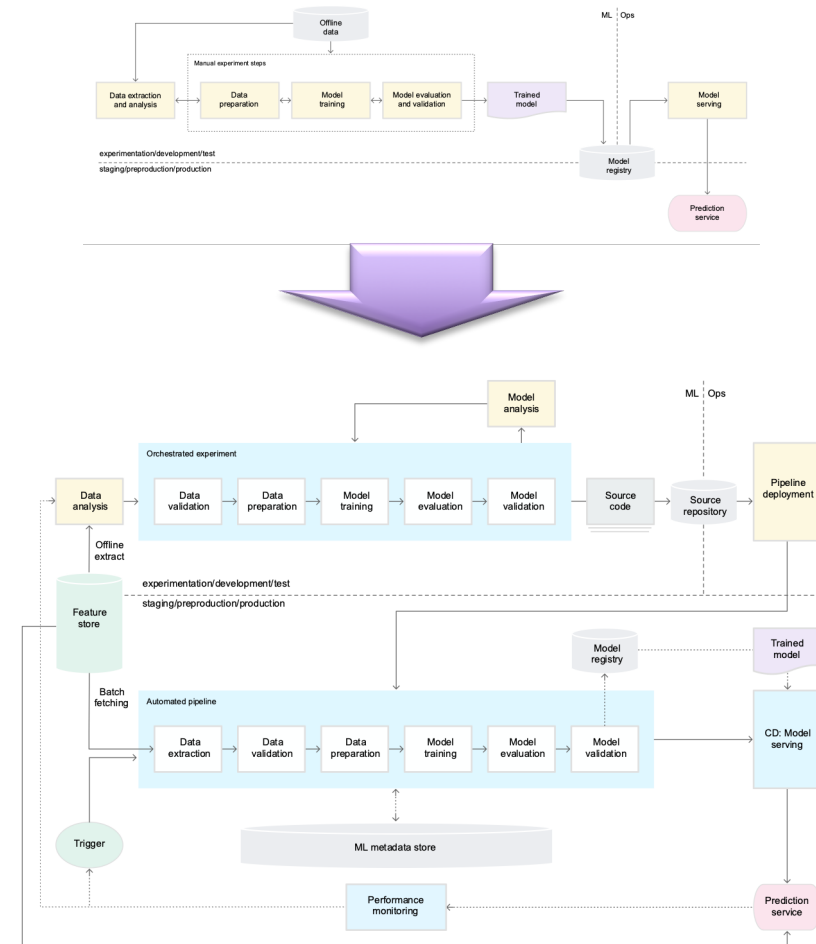
the time for AI/MLOps

- methodologies

- Microsoft's Team Data Science Process
 - <https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview>
- Google's Practitioners guide to MLOps: A framework for continuous delivery and automation of machine learning
 - https://services.google.com/fh/files/misc/practitioners_guide_to_mlops_whitepaper.pdf

- tools

- MLFlow
 - <https://mlflow.org/>
- Kubeflow
 - <https://www.kubeflow.org/>
- Data Version Control & Studio
 - <https://studio.iterative.ai/>



source: <https://cloud.google.com/solutions/machine-learning/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning>

Don't forget

- Curb your enthusiasm...
- A data mining project should always start with an analysis of the data with traditional query tools
 - 80% of the interesting information can be extracted using SQL
 - how many transactions per month include item number 15?
 - show me all the items purchased by Sandy Smith.
 - 20% of hidden information requires more advanced techniques
 - which items are frequently purchased together by my customers?
 - how should I classify my customers in order to decide whether future loan applicants will be given a loan or not?
- Developing and deploying are entirely different beasts!